アソシエーションルール

例題　スーパーマーケットの
商品購買データ

数千の商品

数万
から
数百万
顧客

顧客ID	butter	bread	milk	salad	sugar	…
A	1	0	0	1	0	
B	0	1	1	0	1	
C	0	0	1	0	0	

(1: 購入　0: 未購入)

1顧客平均20個購入すると、1がまばらに出現する表

顧客ID	アイテム集合
A	{butter, salad}
B	{bread, milk, sugar}
C	{milk}
:	:

購入した商品の
集合だけを
保持した表

例題　WEBページに出現するキーワード

URL	アイテム集合 (キーワード集合)
http://www.honda.co.jp/	{automobile, navigation, …}
http://www.nikkei.co.jp/	{stock, economy, …}
http://www.google.com/	{search, engine, …}
:	:

例題　Internet Bookstore での購買履歴

顧客ID	アイテム集合 (本の集合)
A	{complier-text, database-text}
B	{comic, novel}
C	{cooking, gardening}
:	:

例題　遺伝子の組織中での出現データ

組織	アイテム集合（遺伝子mRNA）
小脳	{gene1, gene341, gene562, …}
海馬	{gene3, gene43, gene243, …}
心臓	{gene1, gene2, gene7, gene10, …}
肝臓	{gene3, gene6, gene14, …}
:	:

アソシエーションルールの例

- butter を購入した顧客は milk も購入する確率が高い

- soviet と grandmaster が出現する WEB ページには chess が出現する確率が高い
  （ヒント　IBM Deep Blue vs Kasparov）

- Java の本を購入する顧客は XML の本も購入する確率が高い

- Gene A が出現する細胞では Gene B が出現する確率が高い

用語の定義

- $I = \{i_1, i_2, \ldots, i_m\}$　i_j は記号列でアイテムと呼ぶ
- トランザクションとは I の部分集合であり、唯一の識別子が対応
- D をトランザクションの集合

顧客ID	アイテム集合
A	{butter, salad}
B	{bread, milk, salad}
C	{milk}
:	:

スーパーマーケットの商品購買データ

I = {butter, bread, milk, salad, …}

D = {　{butter, salad},
　　{bread, milk},
　　{milk},
　　:
}

- **アソシエーションルール**とは

$$X \Rightarrow Y$$

の形をしており、X と Y は共通部分がない I の空でない部分集合
（$\phi \neq X \subseteq I$, $\phi \neq Y \subseteq I$, $X \cap Y = \phi$）

例
{butter} ⇒ {milk}
{soviet, grandmaster} ⇒ {chess}
{gene A} ⇒ {gene B}

- I の部分集合 J の**サポート** $\Pr(J)$ とは、
  D の中で J を含むトランザクションの割合
- $X \Rightarrow Y$ の**サポート**とは、$X \cup Y$ のサポート
- $X \Rightarrow Y$ の**確信度**とは、Xを含むトランザクションが Y を含む割合
  すなわち $\Pr(X \cup Y) / \Pr(X)$

D = { {a,b,c,d,e}
  {a,b, d,e}
  {a, c, e}
  {a, e}
  { b,c, e}
  { b, e}
  { c,d,e}
  { d,e} }

Pr({a}) = 4/8=0.5
Pr({b}) = 4/8=0.5
Pr({a,b}) = 2/8=0.25
Pr({a,b,d}) = 2/8=0.25

{a} ⇒ {b} **サポート** 0.25
  確信度 0.5
{a,b}⇒ {d} **サポート** 0.25
  確信度 1

アソシエーションルール の評価基準

$X \Rightarrow Y$ が興味深い (interesting) とは

- サポートが　与えた下限値 min_sup 以上である
  $\Pr(X \cup Y) \geqq \text{min\_sup}$

- 確信度が　与えた下限値 min_conf 以上である
  $\Pr(X \cup Y) / \Pr(X) \geqq \text{min\_conf}$

例　min_sup = 0.25, min_conf = 0.75 のとき

{a} ⇒ {b}　　サポート 0.25 確信度　0.5　は興味深くないが

{a,b}⇒ {d}　サポート 0.25 確信度　1　は興味深い

興味深いアソシエーションルールを枚挙するアルゴリズム Apriori

1. $\Pr(Z) \geqq \text{min\_sup}$ **となる アイテム集合** Z **を枚挙**

2. $Z = X \cup Y$ **となる互いに素で空でない** X **と** Y **について**
   $\Pr(X \cup Y) / \Pr(X) \geqq \text{min\_conf}$
   ならば、 $X \Rightarrow Y$ **は興味深いアソシエーションルールである**
   (注意 $\Pr(X \cup Y) = \Pr(Z) \geqq \text{min\_sup}$ **)**

ステップ1

min_sup = 0.25

Pr({a}) = 0.5

Pr({b}) = 0.5

Pr({a,b}) = 0.25

Pr({a,b,d}) = 0.25

ステップ2

min_conf=0.75

{a} ⇒ {b}　　Pr({a,b}) / Pr({a}) =0.5

{a,b} ⇒ {d}　Pr({a,b,d}) / Pr({a,b}) = 1

ステップ1の実現

$L_k = \{Z \mid Z$ **は** k **個のアイテムを含み**, $\Pr(Z) \geqq \min\_sup\}$

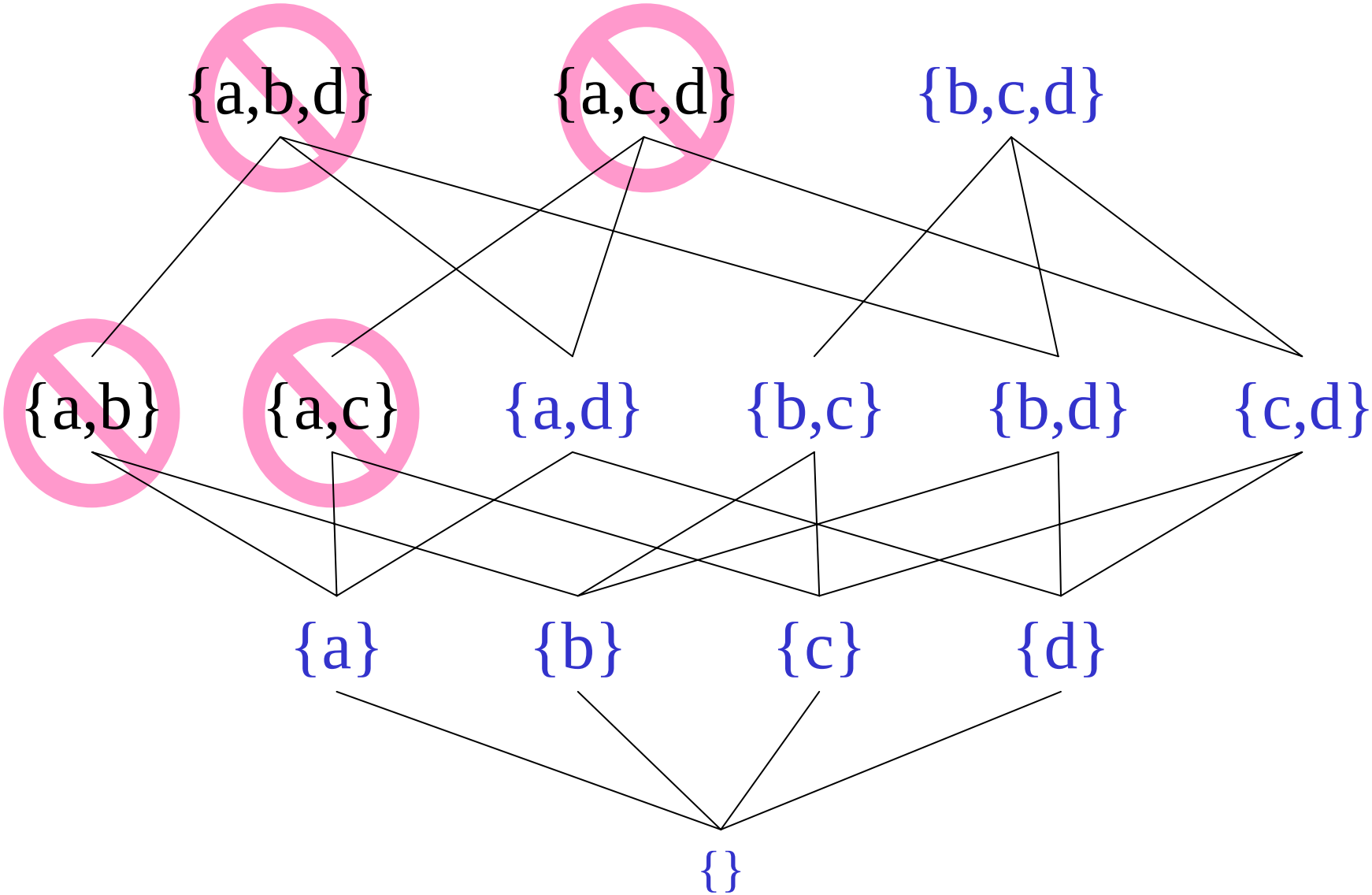
k=1,2,…**の順番に** L_k **を計算し**, $\cup L_k$ **を答えとする**

観察 $L_2 = \{ \{a,d\}, \{b,c\}, \{b,d\}, \{c,d\} \}$

- **単調性** 「$U \subseteq W$ **ならば** $\Pr(U) \geqq \Pr(W)$」 **に注意**
  例 $\Pr(\{a,c\}) \geqq \Pr(\{a,c,d\})$

- $\{a,c\}$ **が** L_2 **の元でないので、**$\min\_sup > \Pr(\{a,c\})$
  すると $\{a,c\}$ **の任意の** superset**、例えば** $\{a,c,d\}$ **も**
  $\min\_sup > \Pr(\{a,c,d\})$
  であり、$\{a,c,d\}$ **は** L_3 **の元ではない**

- L_k **に含まれないアイテム集合の** superset **は**
  L_{k+1} **にも含まれない**


{a,b,d}
{a,c,d}
{b,c,d}
{a,b}
{a,c}
{a,d}
{b,c}
{b,d}
{c,d}
{a}
{b}
{c}
{d}
{}

L_k の元の候補の集合 C_k を、L_{k-1} から生成する手続き

C_k := candidate-generator(L_{k-1}) ;

- **L_k に含まれるアイテム集合の** superset **は L_{k+1} に含まれる可能性がある**

- L_2 = { {a,d}, {b,c}, {b,d}, {c,d} }

  {a,d} **と** {b, d} **の和集合** {a,b,d}　　{a,b} **が L_2 に含まれない**

  {b,c} **と** {b,d} **の和集合** {b,c,d}　　{c,d} **も L_2に含まれる**
  {b,c,d} **は L_3 の元の候補として C_3 に登録**

- k-2 **個のアイテムを共有する L_{k-1} の元**

  $\{i_1, \ldots, i_{k-2}, i_{k-1}\}$ **と**
  $\{i_1, \ldots, i_{k-2}, \quad i_k\}$

  から和集合

  $\{i_1, \ldots, i_{k-2}, i_{k-1}, i_k\}$

  を生成し、k-1 個のアイテムを含む部分集合がすべて L_{k-1} の元 となることを確認したら、C_k に登録

ステップ1の実現

L_1 を求める

for(k:=2; $L_{k\text{-}1} \neq \phi$; k++) **do begin**
　　C_k := candidate-generator($L_{k\text{-}1}$) ;
　　forall $t \in D$ **do begin**
　　　　forall $c \in C_k$ such that $c \subseteq t$ **do begin**
　　　　　　アイテム集合 c に含まれる元の数 $c.count$ を1だけ増やす ($c.count$ ++;)
　　　　end
　　end
　　L_k := { $c \in C_k$ | $c.count$ / |D| ≧ min_sup }
end

return $\cup\{L_i \mid i=1, \ldots, k\text{-}1\}$;

実行例

D = { {1, 3,4}
{ 2,3, 5}
{1,2,3, 5}
{ 2, 5} }

min_sup = 0.5

L_1 = { {1}, {2}, {3}, {5}}

Pr({1})=0.5 Pr({4})=0.25
Pr({2})=Pr({3})=Pr({5})=0.75

C_2 = { {1,2}, {1,3}, {1,5},
{2,3}, {2,5}, {3,5} }

Pr({1,2})=Pr({1,5})=0.25
Pr({1,3})=Pr({2,3})=Pr({3,5})=0.5
Pr({2,5})=0.75

L_2= { {1,3}, {2,3}, {2,5}, {3,5} }

C_3 = { {2,3,5} }

Pr({2,3,5}) = 0.5

L_3 = { {2,3,5} }

- データベースが主記憶に収まらず2次記憶装置に置かれている場合、主記憶と2次記憶間のデータ転送速度がボトルネック

  CPU ⇔ 主記憶 (10^{-8} sec./B)
  主記憶 ⇔ 2次記憶 (≧10^{-2}sec./block, block size 4KB〜56KB)

- Apriori アルゴリズムでは、はじめて $L_k = \phi$ となる k を k^* とするとき、k^* 回だけデータベースを順番に走査する

  現実には k^* は高々 10 程度といわれている

- L_k の大きさは、現実にはデータベースのサイズよりはるかに小さく、主記憶上に格納できる

  しかも C_k により L_k をかなり絞り込むことができる

- 実際のデータベースに適したバランスのとれた実装方法

統計学からの批判： 確信度は意味があるか？

D = {
{a,b,c,d,e}
{a,b, d,e}
{a, c, e}
{a, e}
{ b,c, e}
{ b, e}
{ c,d,e}
{ d,e} }

ルール	サポート	確信度
{a} ⇒ {b}	0.25	0.5
{a,b} ⇒{d}	0.25	1
{a} ⇒ {e}	0.5	1

{a} ⇒ {e} **が最も価値あるルールか？**

Pr({a}) × Pr({e}) = Pr({a,e}) = 0.5

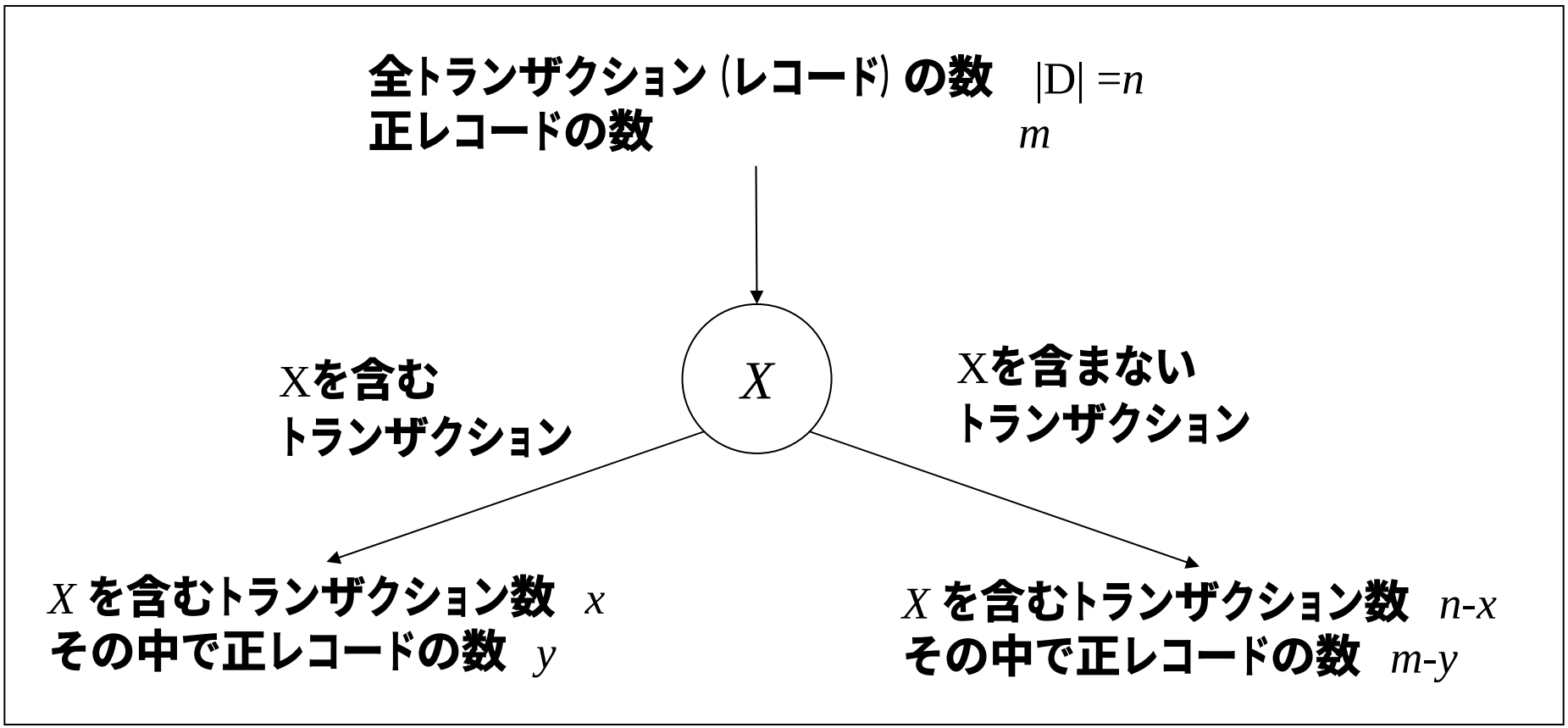
統計学的には {a} **と** {e} **は独立**

ルール $X \Rightarrow Y$ **の評価基準として, 独立性の検定で使うカイ二乗値や、決定木のノードの評価で使った** Entropy Gain **を利用するのが自然という立場がある。 詳しくは**

Shinichi Morishita and Jun Sese. "Traversing Itemset Lattices with Statistical Metric Pruning"
P*roc. of ACM SIGACT-SIGMOD-SIGART Symp. on Database Systems (PODS)*. pp.226-236, May 2000.

決定木とアソシエーションルールの関係

$X \Rightarrow Y$ は深さ1の決定木とみなせる
Y を含むトランザクションを正レコードとみなす



$X \Rightarrow Y$ の確信度 ＝ y / x